STATE LEVEL CROP AREA ESTIMATION USING SATELLITE DATA IN A REGRESSION ESTIMATOR

Mitchell L. Graham, USDA/NASS
3251 Old Lee Hwy. Rm. 305 Fairfax, Virginia 22030

JUNE 1993

KEY WORDS: regression estimator, Landsat Thematic Mapper, land cover estimate, crop acreage estimate.

ABSTRACT

The USDA's National Agricultural Statistics Service (NASS) estimates state level crop acreage in the Mississippi Delta region using area frame survey data and Landsat Thematic Mapper (TM) satellite data. Five general steps produce these acreage estimates. First, a sample of TM pixel data is clustered by land cover. Second, sampled TM pixels are assigned to a land cover class using maximum likelihood classification. Third, classified sample pixels are regressed with reported crop acreages. Fourth, TM scenes are classified. Finally, acreage is estimated with a regression estimator using classified pixel counts as ancillary information to the ground survey data. The potential benefit is mainly a reduction in variance with some adjustment of the state acreage estimates.

BACKGROUND

The Mississippi River Delta region is the most important rice producing area in the United States and is also a major cotton producing area. The region, which includes all or part of five states, accounted for 76 percent of U.S. planted rice acreage and 29 percent of U.S. planted cotton acreage in 1991. With 1.3 million planted acres of rice, Arkansas was the major Delta rice producing state accounting for 46 percent of the 1991 national total. (USDA NASS, 1992). The 1992 Arkansas rice estimate was 1.4 million planted acres; the 1993 estimate was 1.35 million planted acres (USDA NASS, 1993).

The Delta region provides an ideal setting for remote sensing based estimation techniques. NASS's current general purpose area sampling frame is not designed for crops that are localized in specific areas. This condition can lead to high state level relative sampling errors for crops such as cotton and rice. In Arkansas, nearly all the rice and cotton occur in the eastern third of the state oriented north-to-south along the Mississippi River. This geographic orientation coincides with the ground viewing orientation of polar orbiting Landsat satellites and minimizes the number of satellite scenes needed to cover Arkansas.

DATA PROCESSING

PEDITOR is used for data processing on a MicroVax 3500 computer and on IBM PC compatibles in a DOS environment. PEDITOR is a special purpose software system developed at NASS (Ozga et al., 1992) for crop area estimation. PEDITOR is mainly written in PASCAL and contains modules for image display and processing, as well as estimation. Image display and graphics modules are run on PCs, while non-graphics modules can run on a either a PC or MicroVax. Computationally intensive jobs, such as classification of multitemporal TM scenes, are processed on a Cray supercomputer (Idaho National Engineering Laboratory Supercomputing Center in Idaho Falls, Idaho).

DATA ACQUISITION

For the 1991/92 Delta Project, NASS's Remote Sensing Section (RSS) acquired ground data from the June Agricultural Survey (JAS) and Landsat data from EOSAT Corporation. Data acquisition involved the JAS, a recheck visit to JAS segments, spring TM scene selection, and summer TM scene selection.

The ground sample units were small land areas called segments, each about one square mile for strata 11, 12, 20 and 21. Segments were selected randomly from an area sampling frame stratified by land use categories ordered by percent of cultivated land. See Table 1. During the June survey, field enumerators interviewed the land managers in each segment and recorded the land cover (rice, fallow, soybeans, pasture, woods, water, etc.), size, and boundaries for every field. Uncultivated areas within a segment were also recorded. At this point, the survey data could be used to make NASS's usual preliminary crop area estimates having measurable precision, but based on ground data alone. Mid-summer, RSS rechecked segments where a farmer indicated, during the JAS, that a crop would be planted later.

Using knowledge of cropping practices, analysts selected Landsat TM scene dates to facilitate crop discrimination within the constraints imposed by cloud

cover and scene availability. TM data consists of seven spectral measurements on each of 41.6 million picture elements (pixels) arranged in a 5965 by 6967 array called a scene. When possible, spring and summer Landsat TM scenes from the same area were combined to create a single multitemporal, 14 dimensional, satellite data set. Each Landsat scene was reformatted and registered to 1:250,000 USGS maps. Then sampled segments were digitized and located within each Landsat scene. When the geographic correspondence between TM pixel data and JAS segments was established, the Landsat TM data were analyzed by land cover.

Table 1: USDA NASS Land-use Strata for Arkansas during 1991 and 1992.

Stratum #	Definition	n	N
(1991--implemented in 1974)			
11	over 80 % cultivated	144	11,723
12	51 to 80 % cultivated	48	5,697
20	15 to 50 % cultivated	84	11,673
31	agri-urban: > 20 home/mile^2	28	5,019
32	commercial: > 20 home/mile^2	4	1,371
33	resort: > 20 home/mile^2	4	532
40	less than 15 % cultivated	84	10,658
50	non-agricultural	4	889
(1992--implemented in 1992)			
11	over 75 % cultivated	195	11,673
21	25 to 75 % cultivated	40	2,718
31	agri-urban: > 100 home/mile^2	10	1,308
32	commercial: > 100 home/mile^2	5	418
42	less than 25 % cultivated	140	18,561
40	non-agricultural	5	35

Table 2: Landsat TM Scene Overpass Dates for 1991 and 1992 Arkansas Analysis Regions.

Analysis Region	Multi-temporal	Overpass Date Pass 1.	Pass 2.
1991			
Eastern	yes	4/01/91	8/23/91
Central	no	8/14/91	---
1992			
Northeast	yes	5/05/92	7/24/92
Southeast	yes	5/05/92	6/22/92
Central	yes	4/26/92	8/16/92

The TM scene acquisition dates and data quality affect the organization of both analysis and estimation. To control atmospheric and phenological factors, areas of Arkansas viewed by Landsat on different dates are analyzed and processed separately. The Landsat 5 satellite flies North to South over Arkansas in three partially overlapping passes which cover, or view, the eastern, central and eastern regions of the state on different dates. Landsat 5 repeats any given pass every 16 days with neighboring passes either seven or eleven days apart. At best, the central and eastern passes may be seven days apart. In some cases bad weather requires dividing a single path (pass) into two analysis regions that differ by 16, 32 or more days. See Table 2 for TM scene overpass dates.

SATELLITE DATA ANALYSIS

Separately, for each land cover within each analysis region, the segment Landsat data were studied for outlier pixels and then clustered using a modified ISODATA algorithm (Bellow and Ozga, 1991). Outlier pixels were identified using principal component analysis and removed from the data before clustering. The result of clustering each land cover, ç, was several separable vectors, $S_ç$, of spectral reflectance each referred to as a signature. The signatures in $S_ç$ were assumed to represent noticeable variations in the land cover. For example, in S_{rice} separate signatures were expected for unplanted fields, flooded fields, waste areas, fields in good or bad condition, and mixtures of rice and other covers.

When all land covers were clustered, the $S_ç$ were assembled into one collection of signatures, $S_{(all)}$. The separability of the land cover signatures in $S_{(all)}$ was analyzed using Swain-Fu (Swain 1972) or transformed divergence (Swain and Davis 1978) statistics. Some signatures were separable. Most signatures had a degree of separability that would allow them to still be useful for classification. The signatures with the poorest separability were removed from $S_{(all)}$, or averaged with similar signatures, producing an edited collection of signatures, $S_{(edited)}$. Each vector in $S_{(edited)}$ was still tagged with its original land cover but was also considered a separate category of surface reflectance.

Analysts used $S_{(edited)}$ as input into the discriminate function categorizing Landsat TM pixels into separate reflectance categories. There were two phases of maximum likelihood classification. First the segment pixel data were classified. Then after analysis and refinement of segment classification, whole TM scenes were classified.

Analysis of sample segment classification consisted of three parts. First classified segment pixels were tabulated by the reflectance categories in $S_{(edited)}$. Next

commission and omission error based on the original land use tags were examined using the kappa statistic (Congalton, 1991). Then segment classified pixel counts were regressed with segment land cover totals univariately for each land cover. A separate first order model was used in each applicable JAS land use stratum. If classification errors were acceptable and simple linear regression analysis revealed no problems with model assumptions nor outlier points, then the segment classified pixel counts were used to calculate the sample ancillary mean, and b_1 was used to estimate the slope in the regression estimator. Otherwise, some of the satellite data analysis steps were repeated.

When sample level analysis was complete, analysts used $S_{(edited)}$ in classifying whole Landsat scenes. After a TM scene classification, the scene pixels were tabulated within JAS land use strata by category and land cover. These counts were use in calculating the ancillary population means.

REGRESSION ESTIMATOR

Remote sensing researchers at NASS have used ancillary satellite information in a regression estimator since 1978. Analysts used the regression estimator in this manner for land cover and crop estimation projects with the National Aeronautics and Space Administration and the National Oceanic and Atmospheric Administration (Allen and Hanuschak, 1988). There is a theoretical downward bias of order 1/n with this method (Cochran, 1977).

The NASS area frame stratifies each state by percent of cultivated land (Table 1.). Let s = 1,2,...,H denote these land use strata. In each stratum there are N_s primary sampling units (PSU). NASS randomly selects n_s units (segments) from each stratum for enumeration during the JAS.

After purchasing Landsat TM scenes covering the study area, NASS creates analysis regions for the differing satellite overpass dates (Table 2.). Denote the analysis regions $\alpha = 1,2,...,k,k+1,...,A$ where k of them are covered by Landsat data and A-k of them are not.

Within each analysis region, there are H_α area frame land use strata where the regression estimator is used. If the region is covered by Landsat TM data ($\alpha \le k$), $0 \le H_\alpha \le H$. If the region is not covered by TM data ($\alpha > k$), then $H_\alpha = 0$. Denote the area frame land use strata within a covered analysis region as $h = 1,...,H_\alpha$ for strata where the regression estimator is used and as $h = H_\alpha+1,...,H$ for strata where the regression estimator is not used. If the analysis region is not covered by TM data, $h = H_\alpha+1,...,H$.

$$\text{Let } N_s = N_{.h} = \sum_{\alpha=1}^{A} N_{\alpha h} \quad , \qquad \sum_{s=1}^{H} N_s = \sum_{h=1}^{H} N_{.h} \quad ,$$

$$n_s = n_{.h} = \sum_{\alpha=1}^{A} n_{\alpha h} \quad \text{and} \quad \sum_{s=1}^{H} n_s = \sum_{h=1}^{H} n_{.h} \; .$$

The regression estimator of total acreage for a land cover in an analysis region can be expressed as

$$\hat{Y}_{ç\alpha(reg)} = \sum_{h=1}^{H_\alpha} N_{\alpha h} \left[\bar{y}_{ç\alpha h} + \hat{b}_{ç\alpha h}(\bar{X}_{ç\alpha h} - \bar{x}_{ç\alpha h})\right]$$

$$\mathrm{Var}(\hat{Y}_{ç\alpha(reg)}) = \sum_{h=1}^{H_\alpha} (N_{\alpha h}^2 - N_{\alpha h} n_{\alpha h})/n_{\alpha h}$$

$$\cdot \sum_{j=1}^{n_{\alpha h}} (y_{ç\alpha hj} - \bar{y}_{ç\alpha h})^2 (1 - \hat{R}_{ç\alpha h}^2)/(n_{\alpha h} - 2) \cdot [1 + (n_{\alpha h} - 3)^{-1}]$$

Where $b_{ç\alpha h}$ is regression coefficient b_1 for land cover ç region α and stratum h, and where

$\bar{X}_{ç\alpha h} = \sum_{i=1}^{N_{\alpha h}} X_{ç\alpha hi}/N_{\alpha h}$ and $X_{ç\alpha hi}$ is the count of full scene pixels classified to land cover ç in stratum h from the i^{th} PSU in analysis region α.

Likewise, $\bar{x}_{ç\alpha h} = \sum_{j=1}^{n_{\alpha h}} x_{ç\alpha hj}/n_{\alpha h}$ and $x_{ç\alpha hj}$ is the count of segment pixels classified to land cover ç in stratum h from the j^{th} sample unit in analysis region α.

$\hat{R}_{ç\alpha h}^2$ is the coefficient of determination between the reported acreage and classified pixel count of land cover ç for stratum h in analysis region α.

Now for the remaining analysis regions and strata where Landsat TM data were not used, a direct expansion estimator can be expressed as

$$\hat{Y}_{ç\alpha(dir)} = \sum_{h=H\alpha+1}^{H} N_{\alpha h}/n_{\alpha h} \sum_{j=1}^{n_{\alpha h}} y_{ç\alpha hj}$$

$$\mathrm{Var}(\hat{Y}_{ç\alpha(dir)}) = \sum_{h=H\alpha+1}^{H} (N_{\alpha h}^2 - N_{\alpha h} n_{\alpha h})/(n_{\alpha h}^2 - n_{\alpha h}) \sum_{j=1}^{n_{\alpha h}} (y_{ç\alpha hj} - \bar{y}_{ç\alpha h})^2$$

Where $y_{ç\alpha hj}$ is the reported acreage of land cover ç from segment j in stratum h from analysis region α. The state level estimate of land cover ç using ancillary Landsat TM data is written

$$\hat{Y}_{TMç} = \sum_{\alpha=1}^{k} \hat{Y}_{ç\alpha(reg)} + \sum_{\alpha=1}^{k} \hat{Y}_{ç\alpha(dir)} + \sum_{\alpha=k+1}^{A} \hat{Y}_{ç\alpha(dir)}$$

$$Var(\hat{Y}_{TMç}) = \sum_{\alpha=1}^{k} Var(\hat{Y}_{ç\alpha(reg)}) + \sum_{\alpha=1}^{k} Var(\hat{Y}_{ç\alpha(dir)}) + \sum_{\alpha=1}^{A} Var(\hat{Y}_{ç\alpha(dir)})$$

RESULTS

For 1991 and 1992, the Remote Sensing Section submitted Landsat crop acreage indications to the NASS Agricultural Statistics Board and the Arkansas State Statistical Office early in December. NASS's Annual Crop Production Report, published in early January, contained crop acreages from the December board.

Before submission, the acreage indications are assessed through examining statistics from each of the main processing steps. Classification accuracy, exclusion error, and inclusion error are assessed using the kappa statistic, percent correct and percent commission. The regression relationship of acres with classified pixels is analyzed for fit, outlier segments and appropriate slope. Since the Landsat TM pixel is approximately 0.201 acres, then b_1 should be near 0.201. Also, the relative efficiency (RE) of the state level Landsat regression estimator to that of the direct expansion (JAS) estimate is noted.

Table 3 gives the kappa statistic, and percent correct and percent commission for rice in Arkansas for 1991 and 1992. Commission errors were better in 1992 with substantially better classification accuracy for 1992 central region.

For both 1991 and 1992 the central and eastern areas of Arkansas were covered by TM scenes. Weather conditions in each year were the final determinate for TM scene selection. In 1991 acceptable TM data were obtained only for mid-summer over the central analysis region while early spring and mid-summer data were available for the eastern region. Consequently, the 1991 central region was analyzed with unitemporal TM data while the eastern region was multitemporal. In 1992, spring as well as summer imagery was available, so that multitemporal TM data sets were created for all regression analysis regions. But the 1992 eastern region had differing summer image dates for northeast and southeast and was therefore divided into two analysis regions to control for atmospheric and crop progress effects. In general, classification accuracy was higher in the multitemporal analysis regions than in the unitemporal regions.

Table 4 shows the stratum level sample sizes ($n_{\alpha h}$) and $R_{ç\alpha h}^2$ values for those strata where regression was used for rice. Table 5 shows state level direct expansion CV's (CV_{DE}), Landsat regression CV's (CV_{TM}), and the RE's for rice. Table 6 shows the difference of total planted rice acres estimated by direct expansion only from the estimate produced through using the regression estimator scaled by standard error. The state level and analysis region acreage indications (unofficial estimates) cannot be shown due to confidentiality restrictions.

In 1991, both state level direct expansion and regression method indications for planted acres of rice were below the 1991 official NASS estimate, while for 1992 the official estimate was between these two 1992 indications. In 1991, $\hat{Y}_{DE}$ was closer to the official estimate, but in 1992 $\hat{Y}_{TM}$ differed very little from the official NASS estimate. $\hat{Y}_{TM,1991}$ was 1.28 standard errors ($SE_{TM,1991}$) below the 1991 official rice estimate, and $\hat{Y}_{TM,1992}$ was 0.53 standard errors ($SE_{TM,1992}$) above the 1992 official estimate.

Table 3: Kappa values (k), percent correct (ct) and percent commission (cm) for sample segments' classification -- All Rice.

	Analysis Regions								
	Northeast[1]			Southeast[1]			Central[2]		
Cover	k	ct	cm	k	ct	cm	k	ct	cm
rice (1991)	71	75	27				67	68	27
rice (1992)	74	79	19	81	84	14	83	87	14

Table 4: Regression of Reported Segment Acreage with Segment Categorized Pixels for All Rice.

	Analysis Regions								
	Northeast[1]			Southeast[1]			Central[2]		
Stratum[5]	n	R^2	b	n	R^2	b	n	R^2	b
1991[3]									
11	98	.94	.194	---[1]			23	.96	.222
12	13	.99	.204	---[1]			9	---[4]	
1992[3]									
11	54	.95	.195	53	.98	.203	37	.98	.191
21	1	---[4]		10	.84	.174	7	.97	.190

Table 5: Arkansas State Level Relative Efficiency (RE) for All Rice.

Crop	CV_{DE}(%)	CV_{TM}(%)	RE
Rice (1991)	10.1	5.4	3.9
Rice (1992)	6.8	4.1	3.2

Table 6: Difference in Total Planted Acreage. Direct Expansion Estimate minus Regression Method Estimate Scaled by Standard Error.

Crop	$(\hat{Y}_{DE}-\hat{Y}_{TM})/SE_{DE}$	$(\hat{Y}_{DE}-\hat{Y}_{TM})/SE_{TM}$
Rice (1991)	0.50	0.99
Rice (1992)	1.32	2.36

SUMMARY

In 1991 and 1992, the NASS Remote Sensing Section estimated planted rice acreage in Arkansas using NASS June Agricultural Survey area frame data and ancillary Landsat TM data in a regression estimator. To control for phenological effects, Arkansas was divided into analysis regions based on TM scene overpass dates. Each analysis region was analyzed separately. A regression estimator was used within the intensively cultivated land use strata for the TM covered analysis regions; otherwise, direct expansion was used. The state level acreage estimate was the sum of the analysis region estimates. For 1991, the regression estimator produced a state level indication (unofficial estimate) which was 1.28 standard errors below the NASS official planted acres estimate for rice. In 1992, the indication was 0.53 standard errors above the official estimate. For each year, the regression method indication and variance were less than the corresponding direct expansion indication and variance.

---

[1] The northeast and southeast regions were analyzed as one region in 1991 and as two in 1992.

[2] The central region was analyzed unitemporally in 1991.

[3] The Arkansas area sampling frame was reconstructed for 1992.

[4] Direct expansion was used.

[5] Direct expansion was used in strata which are not listed.

[DE] Direct expansion method--no ancillary satellite data used.

[TM] Method using regression estimator with satellite data where possible and direct expansion where not.

ACKNOWLEDGEMENTS

I would like to thank Paul Cook of USDA NASS for his 1991 analysis of central Arkansas.

REFERENCES

Allen, J.D., 1990a, A Look at the Remote Sensing Applications Program of the National Agricultural Statistics Service, Journal of Official Statistics, 6(4):393-409.

Allen, J.D., 1990b, Remote Sensor Comparison for Crop Area Estimation Using Multitemporal Data, in Proceedings of the IGARSS '90 Symposium, College Park, Md., pp. 609-612.

Allen, J.D. and Hanuschak, G.A., 1988, The Remote Sensing Applications Program of the National Agricultural Statistics Service: 1980-1987, U.S. Department of Agriculture, NASS Staff Report No. SRB-88-08.

Bellow, M.E. and Graham, M.L., 1992, Improved Crop Area Estimation in the Mississippi Delta Region Using Landsat TM Data, in Proceedings of the ASPRS/ACSM Convention, Washington, D.C.

Bellow, M.E. and Ozga, M., 1991, Evaluation of Clustering Techniques for Crop Area Estimation using Remotely Sensed Data, in American Statistical Association 1991 Proceedings of the Section on Survey Research Methods, Atlanta, Ga., pp. 466-471.

Cochran, W.G., 1977, Sampling Techniques, John Wiley and Sons, New York, NY, ch. 7, pp 189-203.

Cook, P.W., 1982, Landsat Registration Methodology Used by U.S. Department of Agriculture's Statistical Reporting Service 1972-1982, USDA/NASS/Remote Sensing Section.

Congalton, R.G., 1991, A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data, Remote Sensing of the Environment, 37:35-46 (1991).

Cotter, J. and Nealon, J., 1987, Area Frame Design for Agricultural Surveys, U.S. Department of Agriculture, NASS Area Frame Section.

Gong, P. and Howarth, P.J., 1992, Frequency-Based Contextual Classification and Grey-Level Vector Reduction for Land-Use Identification, Photogrammetric Engineering and Remote Sensing, Vol. 58, No. 4, April 1992, pp 423-437.

Johnson, R.A. and Wichern, D.W., 1988, Applied Multivariate Statistical Analysis, Prentice Hall, Englewood Cliffs, N.J., ch. 11, pp. 501-513.

Ozga, M., Mason, W.W. and Craig, M.E., 1992, PEDITOR - Current Status and Improvements, in Proceedings of the ASPRS/ACSM Convention, Washington, D.C.

U.S. Department of Agriculture, 1992, Crop Production - 1991 Summary, Agricultural Statistics Board, NASS.